

Structure of Heterogeneous Networks

Rumi Ghosh

University of Southern California
Information Sciences Institute
Marina del Rey, California 90292
Email: rumi.ghosh@gmail.com

Kristina Lerman

University of Southern California
Information Sciences Institute
Marina del Rey, California 90292
Email: lerman@isi.edu

Abstract—Heterogeneous networks play a key role in the evolution of communities and the decisions individuals make. These networks link different types of entities, for example, people and the events they attend. Network analysis algorithms usually project such networks unto simple graphs composed of entities of a single type. In the process, they conflate relations between entities of different types and lose important structural information. We develop a mathematical framework that can be used to compactly represent and analyze heterogeneous networks that combine multiple entity and link types. We generalize Bonacich centrality, which measures connectivity between nodes by the number of paths between them, to heterogeneous networks and use this measure to study network structure. Specifically, we extend the popular modularity-maximization method for community detection to use this centrality metric. We also rank nodes based on their connectivity to other nodes. One advantage of this centrality metric is that it has a tunable parameter we can use to set the length scale of interactions. By studying how rankings change with this parameter allows us to identify important nodes in the network. We apply the proposed method to analyze the structure of several heterogeneous networks. We show that exploiting additional sources of evidence corresponding to links between, as well as among, different entity types yields new insights into network structure.

I. INTRODUCTION

Heterogeneous networks play a key role in information dissemination, evolution of communities, and the decisions individuals make. While traditional network analysis algorithms can efficiently find structure even in large data sets, they usually work on homogeneous data, i.e., networks composed of entities of a single type, for example, a social network where individuals are nodes and an edge between nodes corresponds to a (possibly directed) friendship relationship. Such networks can be represented as unipartite graphs. Many online networks, however, mix entities of different types. On the popular photo-sharing site Flickr, for example, users can post images, tag them with descriptive keywords, join special-interest photography groups, befriend other users, mark images of others users as their favorite, and so on. We can represent Flickr as a *heterogeneous* network composed of several entity types: *users*, *images*, *groups*, and *tags*, with connections between the entities representing different types of relations. A link between users denotes a friendship; a link between a user and a group denotes user's membership in the group; a link between an image and tags represents the keywords used to annotate that image, and so on. In order to extract useful knowledge from this data, we need to look at the network in its entirety.

Heterogeneous networks are sometimes represented as multi-partite (bipartite, etc.) graphs in which vertices are partitioned into disjoint sets corresponding to different entities, with edges connecting vertices from different sets. While some approaches consider the bipartite graph as the whole [7], others usually project the network onto a unipartite graph for matrix algebraic analysis [30]. Thus, the user-group network is reduced to a graph containing users only, with links between users denoting membership in the same photo group. Such projections, however, lose important information about network structure [1]. Recently, [1], [2] extended the modularity-based approach [3] to find structure in bipartite graphs. They showed that compared to analyzing a projected graph, taking into account links between different entity types leads to better community structure. However, bipartite graphs do not allow for edges between vertices of the same type; thus, methods based on this representation cannot exploit all the information available in a heterogeneous network. In the user-group network, for example, information from friendship links between users could augment information from group membership, leading to a better understanding of the network.

This paper makes two contributions. First, in Section II, we present a mathematical framework for compactly representing a heterogeneous network that combines entities and links of different types. We represent such a network as a multi-layer graph, where each layer contains vertices (entities) of a unique type, with edges linking vertices across different layers, as well as within a single layer. Thus, the user-group network is a 2-layer graph, with intra-layer edges in the user layer giving user-user (friendship) relations, and inter-layer edges giving the user-group (membership) relations. Using this mathematical representation, we develop algorithms to study network structure. We use Bonacich centrality [4] as the basis for network analysis, specifically for identifying communities and important nodes in a network. This centrality metric, defined in Section III, gives the number of paths of any length linking two nodes. It contains a tunable parameter that allows us to set the length scale of the interactions. As the second contribution of the paper, we extend the modularity-based community detection algorithm to find groups of nodes that are more connected, in Bonacich centrality sense, to each other than to outside nodes. We also use Bonacich centrality to rank individual nodes in the network. Finally, in Section IV, we apply this framework to study the structure of real-world

heterogeneous networks. We analyze two benchmark networks studied in literature, as well as a network extracted from the social photosharing site Flickr. We show that exploiting information contained in links *between*, and *among*, different entity types leads to new insights into network structure.

II. N-MODE MATRIX REPRESENTATION

We compactly represent a heterogeneous network as a layered graph, in which entities belonging to different classes are partitioned into separate layers, with *intra-layer* and *inter-layer* edges representing links between entities. Consider a network with two entity classes X ($|X| = n$) and Y ($|Y| = m$). For concreteness, suppose the data represents a scientific papers dataset with authors X and papers Y , and that in addition to the usual authorship relations, we managed to collect additional data about friendships, acknowledgements and citations. This data can be represented as a graph with two layers, with vertices of type X (authors) in one layer, and vertices of type Y (papers) in the other layer. An $(m+n) \times (m+n)$ adjacency matrix captures the intra- and inter-layer relations between different vertices:

$$A = \begin{bmatrix} XX_{m \times m} & XY_{m \times n} \\ YX_{n \times m} & YY_{n \times n} \end{bmatrix}$$

Here $A_{ij} = XX_{ij}$ gives the binary relation of the ordered pair (x_i, x_j) , e.g., a friendship between authors i and j ; $A_{i,j+m} = XY_{ij}$ gives the binary relation of the ordered pair (x_i, y_j) , e.g., if author i wrote paper j ; $A_{i+m,j} = YX_{ij}$ gives the binary relation of the ordered pair (y_i, x_j) , e.g., if paper i acknowledges author j ; $A_{i+m,j+m} = YY_{ij}$ gives the binary relation of the ordered pair (y_i, y_j) , e.g., whether paper i cites paper j . We call this data structure a *2-mode matrix*. This representation is similar to one used by Tong et al. [5] to represent bipartite graphs, except since bipartite graphs only describe the inter-layer, and not the intra-layer, relations, the diagonal submatrices XX and YY are zero.

We can easily generalize the above formulation to N -mode matrices, which represent graphs having N distinct types of nodes or being composed of entities belonging to N distinct classes. The adjacency matrix in this case represents N^2 distinct types of binary relations. Now that we have a mathematical representation of heterogeneous networks, we are ready to explore their structure.

III. NETWORK CENTRALITY AND STRUCTURE

Centrality measures the degree to which network structure determines importance of a node in a network. Social network researchers have proposed several different measures of centrality [4], [6], [7] to explain the influence or status of individual actors within a social network. Katz [6], for example, recognized that an individual actor's centrality depends not only on how many others she is connected to (her degree), but also on the centrality of the players she is connected to. Katz score measures status of an actor by the total number of paths linking it to other nodes in the network, exponentially weighted by the length of the path [6]. Bonacich [4] generalized this idea by introducing a new measure of centrality,

$C(\alpha, \beta)$, parameterized by α and β . Bonacich centrality (*b-centrality*) measures the expected number of transmissions directly or indirectly caused by a node. Like Katz score, b-centrality is given by the total number of attenuated paths emanating from a node, but now the attenuation factors along direct links, β , and indirect links, α , in a path can be different¹

$$\begin{aligned} C(\alpha, \beta) &= (\beta A + \beta \alpha A \cdot A + \dots + \beta \alpha^n A^{n+1} \dots) \\ &= \beta A(I - \alpha A)^{-1}. \end{aligned} \quad (1)$$

This equation holds while $\alpha < 1/\lambda$, where λ is the largest characteristic root of A [8]. For $\alpha = \beta$, this measure reduces to the Katz status score [6]. B-centrality can be easily generalized to heterogeneous networks, with matrix A corresponding to the N -mode matrix representing the network. We can use b-centrality to study the structure of a heterogeneous network. We extend the popular modularity-based community detection method to utilize b-centrality. In addition, we show that b-centrality can identify influential nodes within the network, as well as nodes that bridge different communities. One advantage of using b-centrality is that we can vary parameter α to set the length scale of the interactions. α is the probability of transmitting a message or influence along an indirect edge in a path emanating from a vertex. The expected length of a path, the radius of centrality, is $(1 - \alpha)^{-1}$. For $\alpha = 0$, b-centrality takes into account direct edges only. Many network analysis algorithms use such local structures, e.g., the degree of the vertex, as the metric in their analysis. As α increases, b-centrality becomes a more global measure, taking into account ever larger network components. This tunable parameter turns b-centrality into a powerful tool for investigating network structure. In real-world networks, we can estimate the value of α along a particular link of a network by measuring the probability that a node transmits a message received from a distant node along this link. In most situations, however, this information is not readily available. Although we may not know its exact value, studying how network properties change with α gives us valuable insight into network structure.

A. Community Detection

Girvan & Newman [3] proposed *modularity* as a metric for evaluating community structure of a network. The modularity-optimization class of community detection algorithms [9]–[11] find a network division that maximizes the modularity Q , given by $Q = (\text{connectivity within community}) - (\text{expected connectivity})$, where connectivity is density of edges. We extend this approach and use b-centrality as the measure of network connectivity [12]. Therefore, in the best division of the network, nodes have more paths connecting them to nodes within their community than to outside nodes. We generalize modularity Q as

$$Q(\alpha) = \sum_{ij} [C_{ij} - \bar{C}_{ij}] \delta(s_i, s_j) \quad (2)$$

¹For some types of networks, e.g., commodity exchange networks, Bonacich allows $\alpha < 0$. In communication and information networks we are considering, $\alpha > 0$. Also, in this paper we reverse Bonacich's notation and take β as direct attenuation and α as indirect attenuation factors.

where C_{ij} is given by Eq. 1, \bar{C} is the expected b-centrality, and s_i is the index of the community i belongs to, with $\delta(s_i, s_j) = 1$ if $s_i = s_j$; otherwise, $\delta(s_i, s_j) = 0$. We round the values of C_{ij} to the nearest integer. Since β factors out of modularity, we consider dependence on α only. To compute the expected centrality, we consider a graph, referred to as the null model, which has the same number of vertices and edges as the original graph, but in which the edges are placed at random. To make the derivation below more intuitive, instead of b-centrality we talk of the number of paths. When all the vertices are placed in a single group, then axiomatically, $Q = 0$. Therefore $\sum_{ij} [C_{ij} - \bar{C}_{ij}] = 0$, and we set $W = \sum_{ij} \bar{C}_{ij} = \sum_{ij} C_{ij}$. Therefore, according to the argument above, the total number of paths between vertices in the null model $\sum_{ij} \bar{C}_{ij}$ is equal to the total number of paths in the original graph, $\sum_{ij} C_{ij}$. We further restrict the choice of null model to one where the expected number of paths reaching vertex j , W_j^{in} , is equal to the actual number of paths reaching the corresponding vertex in the original graph. $W_j^{in} = \sum_i \bar{C}_{ij} = \sum_i C_{ij}$. Similarly, we also assume that in the null model, the expected number of paths originating at vertex i , W_i^{out} , is equal to the actual number of paths originating at the corresponding vertex in the original graph $W_i^{out} = \sum_j \bar{C}_{ij} = \sum_j C_{ij}$. Next, we reduce the original graph G to a new graph G' that has the same number of vertices as G and total number of edges W , such that each edge has weight 1 and the number of edges between nodes i and j in G' is C_{ij} . Now the expected number of paths between i and j in graph G could be taken as the expected number of the edges between vertices i and j in graph G' , and the actual number of paths between vertices i and j in graph G can be taken as the actual number of edges between vertex i and vertex j in graph G' . The equivalent random graph G'' is used to find the *expected* number of edges from vertex i to vertex j . In this graph the edges are placed in random subject to constraints:

- The total number of edges in G'' is W .
- The out-degree of vertex i in G'' = out-degree of vertex i in $G' = W_i^{out}$.
- The in-degree of a vertex j in graph G'' = in-degree of vertex j in graph $G' = W_j^{in}$.

Thus in G'' the probability that an edge will emanate from a particular vertex depends only on the out-degree of that vertex; the probability that an edge is incident on a particular vertex depends only on the in-degree of that vertex; and the probabilities of the two vertices being the two ends of a single edge are independent of each other. In this case, the probability that an edge exists from i to j is given by $C(\text{emanates from } i) \cdot C(\text{incident on } j) = (W_i^{out}/W)(W_j^{in}/W)$. Since the total number of edges is W in G'' , therefore the expected number of edges between i and j is $W \cdot (W_i^{out}/W)(W_j^{in}/W) = \bar{C}_{ij}$, the expected the expected b-centrality in G . Once we compute modularity $Q(\alpha)$ for the N-mode matrix representing the network, we have to select an algorithm to divide the network into communities that optimize $Q(\alpha)$. Brandes et al. [13] have

shown that the decision version of modularity maximization is NP-complete. Like others [11], [14], we use the leading eigenvector method to obtain an approximate solution. In this method, vertices are assigned to either of two groups based on a single eigenvector corresponding to the largest positive eigenvalue of the modularity matrix (spectral optimization of modularity).

As the network grows, matrix A may become quite large, making computation of inverse in Eq. Equation 1 expensive. We use an approximation method, along the lines of [15], that keeps the first three terms in Eq. Equation 1 only.

B. Node Ranking

Social scientists have long believed that structure of the network can affect an individual's productivity and success [16], [17] and predict new links [18] (or ties). Much of the analysis done by social scientists considered local structure, i.e., the nature of an individual's ties [16], [17], [19].

By focusing on local structure, the traditional microscopic theories fail to capture the global, macroscopic structure of the network. This structure is better captured by metrics that take into account *paths* and not merely *links* or ties between nodes. Several different centrality metrics take paths into account to identify nodes that are 'close' in some sense to other nodes in the network, and are therefore, more important. Betweenness centrality [7] calculates node's score as the ratio of the number of shortest paths via the given node to the number of shortest paths in the network. As described above, Katz score [6] of node i is the sum over all paths from i , exponentially weighted by the length of the path. PageRank [20], roughly, gives the probability that a random walk initiated at node i will reach j . Liben-Nowell and Kleinberg [18] evaluated performance of the different scoring mechanisms on the link prediction task and showed that Katz score is one of the most effective measures for this task.

We follow Bonacich [4] and use b-centrality $C(\alpha, \beta)$ as the measure of proximity between nodes in a network. As mentioned above, this metric is a generalization of the Katz score, and enables us to identify important nodes in the network. A node could have a high b-centrality if it is connected to many nodes within its community — these are community *leaders*. A node could also have a high b-centrality if it is connected to nodes in different communities. While they may be peripheral to any given community, these nodes play an important *bridging* role in the network: they mediate communication between communities. We can identify such nodes, because their b-centrality increases as α , the weight of distant links, grows. Other centrality metrics do not distinguish between leaders and bridges.

IV. EMPIRICAL RESULTS

We apply the formalism developed above to study the structure of two heterogeneous networks that have been studied in literature, the College Football [21] and the Southern Women [22] datasets datasets, as well as the user-group network extracted from the social photosharing site Flickr.

We adopt normalized mutual information, MI , as the metric for evaluating the quality of discovered communities [2], [23]. Suppose our method found a community division X , whereas the actual community division of the network is Y . The probability that a node is assigned to group x by the algorithm, whereas it actually belongs to group y is $P(x, y) = N_{xy}/n$, where N_{xy} is the number of nodes that were assigned to x that belong to group y , and n is the total number of nodes. Following Barber [2], we express normalized mutual information as

$$MI(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)},$$

where standard mutual information and entropy are defined as $I(X, Y) = \sum_{x, y} P(X, Y) \log \frac{P(X, Y)}{P(X)P(Y)}$, $H(X) = \sum_x P(X) \log P(X)$, and $H(Y) = \sum_y P(Y) \log P(Y)$. When $MI = 1$, the discovered groups are the actual communities in the network. When $MI = 0$, the discovered groups are independent of the actual communities.

A. Southern Women

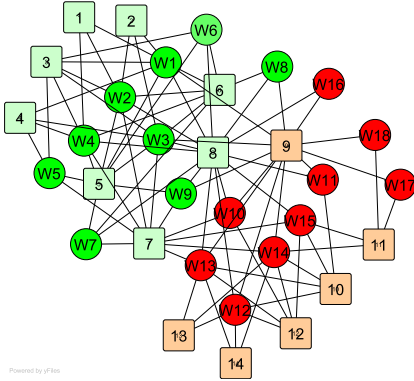


Fig. 1. Bipartite graph representing the Southern Women dataset. Circles represent women and squares the events they attended. Nodes in red and pink belong to one group, while those in green and light green to the other.

The Southern Women dataset comes from a comparative study of social class by Davis et al. [24]. The researchers collected systematic data on the social activities of 18 women over a nine month period. During this time, various subsets of women met in a series of 14 informal events, as shown in Figure 1. Many researchers have subsequently tried to predict the social classes and the structure of groups in this dataset [31]. Freeman [22] reviewed 21 such studies and performed a meta-analysis of predictions to find the groups. He did a canonical analysis of symmetry and dynamic-paired-comparison scaling to find women’s positions (rank) within the group. We take Freeman’s meta-analysis as ground truth for our study.

1) *Communities*: We set $\alpha \leq 0.16$, the reciprocal of the largest eigenvalue of the 2-mode matrix. For all values of α women w1 – w9 were assigned to Group1, and w10 – w18 to Group2, the same results as the ground truth found by Freeman’s meta-analysis [22]. Events 1 to 8 were assigned

to Group1, and events 9 to 14 Group2. The mutual information metric was $MI = 1$. Only 6 of the 21 algorithms in Freeman meta-analysis replicated the ground truth.

Alternatively, the bipartite women-events network can be projected onto a unipartite graph of women only, where a link between women exists if they attended an event together. The community detection algorithm put w2, w4–w7 in one group and the rest of the women in the other, resulting in $MI = 0.38$. As Guimera [1] noted, such projections loose information contained in the bipartite graph. A projection that results in a weighted adjacency matrix, i.e., where A_{ij} shows the number of events women i and j attended together, preserves this information. For $\alpha \leq 0.01$, the reciprocal of the largest eigenvalue of the unipartite matrix, the community detection results agreed with the ground truth, resulting in $MI = 1$.

2) *Rankings*: When we ranked the women we obtained interesting insights into the structure of groups. Table I shows the rankings of women within each group, with 1 as the highest rank. When only the direct links are considered ($\alpha = 0$), then in Group1, w1, w2, w3, and w4 form the core, w5, w6, w7 the primary and w8, w9 the secondary members (with w9 ranking higher than w8), as predicted originally by Davis et al. [24].² However, when we increase α , a crisper ordering emerges. Among the core members, w3 takes the leadership position, followed by w1, w4 and w2. More interestingly, as the strength of indirect links grows, the importance of peripheral members changes. For $\alpha = 0.1$, w9 is ranked higher than w5, and she keeps moving up in rank with increasing α . Woman w9 is a peripheral member of Group1, and is also connected to Group2. In fact the original study assigns w9 as a secondary member of both groups, because she was “claimed” by both groups [24]. Such peripheral members loosely connected to both communities act as bridges between communities and are responsible for the spread of information from one community to another [17], [19]. Our analysis easily identifies these important people.

In Group2, w14 emerges as the leader for $\alpha = 0$, followed by w13 and w12. Women w15 and w11, who are given the same rank, come next, followed by w10, w17, and w18, all with the same rank. Woman 16 has the lowest rank. On increasing α , w16 becomes increasingly more important, surpassing w17 and w18, i.e., she emerges as the “bridge”. Woman w16 is connected to only one node in Group2, yet she becomes more important than w17 and w18 (both connected to two nodes in the group), because she is also connected to Group1. Much the same way, the gradual increase in the ranking of w10 with α can be attributed to her connection to the other group. When $\alpha = 0.12$, there is a change in ranking of w13 and w14, with w13 taking the leadership position. This is the ranking obtained in the meta-analysis, with only w16 ranking above w17 and w18 by our method as compared to the meta-analysis. At $\alpha = 0.14$ w10 ranks above w11. It is, in

²Core, primary and secondary was used in [24] to assign the status of the women within the group.

ID	Meta1	Meta2	$\alpha = 0$	$\alpha = 0.02$	$\alpha = 0.04$	$\alpha = 0.06$	$\alpha = 0.08$	$\alpha = 0.1$	$\alpha = 0.12$	$\alpha = 0.14$	$\alpha = 0.16$
w1	1	1	2.5	2	2	2	2	2	2	2	2
w2	2	2	2.5	4	4	4	4	4	4	4	4
w3	3	3	2.5	1	1	1	1	1	1	1	1
w4	4	4	2.5	3	3	3	3	3	3	3	3
w5	5.5	5	6	7	7	7	7	8	8	8	8
w6	5.5	6	6	5.5	6	6	6	6	6	7	7
w7	7	7	6	5.5	5	5	5	5	5	5	6
w8	9	9	9	9	9	9	9	9	9	9	9
w9	8	8	8	8	8	8	8	7	7	6	5
w10	6	6	7	6	6	6	6	6	6	5	4
w11	5	5	4.5	4.5	4.5	5	5	5	5	6	6
w12	3	3	2.5	3	3	3	3	3	3	3	2
w13	1	1	2.5	2	2	2	2	2	1	1	1
w14	2	2	1	1	1	1	1	1	2	2	3
w15	4	4	4.5	4.5	4.5	4	4	4	4	4	5
w16	9	9	9	9	9	9	9	9	7	7	7
w17	7.5	7.5	7	7.5	7.5	7.5	7.5	7.5	8.5	8.5	8.5
w18	7.5	7.5	7	7.5	7.5	7.5	7.5	7.5	8.5	8.5	8.5

TABLE I
RANKINGS OF WOMEN WITHIN THEIR GROUPS IN THE SOUTHERN WOMEN DATASET. META1 REFERS TO CANONICAL ANALYSIS, AND META2 TO PAIRED COMPARISON. THE REMAINING COLUMNS GIVE RANKINGS FOR DIFFERENT VALUES OF α .

fact, the two women (w10 and w16), who do not conform to the ranking of the meta analysis, who act as bridges between communities.

We applied the same analysis to the unipartite graph, which is the projection of the bipartite data unto a graph of women only. The rankings of Group1 women were somewhat different from the results of the meta-analysis. Whereas the meta-analysis assigned the highest rank to w1, in our analysis w3 had that position. The rankings were mostly independent of α , with only the rankings of w6 and w7 changing as α increased. The rankings of Group2 women were also almost independent of α . At $\alpha = 0$, w11 and w15 had the same rank, but when α grew, w15 was higher ranked. The rankings are similar to the ground truth, with the only difference being w16, who is placed above w17 and w18 by our algorithm, but below w17 and w18 in the rankings obtained from the meta-analysis. In summary, although b-centrality-based rankings produced by the bipartite and unipartite methods were similar to the ground truth, bipartite method allowed us to identify “bridges” who facilitate communication between different communities.

B. College Football

The US College football dataset [21] represents the schedule of Division 1 games for the 2001 college football season. The teams are divided into *conferences* containing 8 to 12 teams each. Games are more frequent between members of the same conference. Inter-conference games, however, are not uniformly distributed, with teams that are geographically closer likely to play more games with one another than teams separated by geographic distances.

We represent the College Football dataset as a 2-mode matrix. The games between teams give the team-to-team relations, while the conferences to which they belong give the team-to-conference relations. Unlike the Southern Women dataset, a purely bipartite network, this dataset contains both the relations among teams and between teams and conferences.

College Football				Flickr			
(a) unimodal		(b) bimodal		(c) unimodal		(d) bimodal	
α	MI	α	MI	α	MI	α	MI
0	0.695	0	0.719	0	0.288	0	0.1682
0.02	0.699	0.005	0.719	0.001	0.286	0.002	0.16
0.04	0.699	0.01	0.719	0.002	0.288	0.004	0.148
0.06	0.658	0.015	0.719	0.003	0.289	0.006	0.083
0.08	0.658	0.02	0.719	0.004	0.289	0.008	0.110
0.1	0.711	0.025	0.719	0.005	0.290		
0.12	0.682	0.03	0.732	0.006	0.292		
0.14	0.684			0.007	0.293		
0.16	0.684			0.008	0.294		
0.18	0.557			0.009	0.296		
				0.01	0.271		

TABLE II
NORMALIZED MUTUAL INFORMATION MEASURE OF COMMUNITIES DISCOVERED IN THE NETWORKS. UNIMODAL REFERS TO NETWORKS CONTAINING NODES OF A SINGLE TYPE, WHILE BIMODAL REFERS TO HETEROGENEOUS NETWORKS WITH TWO TYPES OF NODES.

We used modularity-based approach to find communities for different values of α , $0 \leq \alpha \leq 0.03$. We find eight groups independent of α ; however, setting α to its maximum value leads to purer groups. Table II(b) shows the mutual information-based measure of the quality of the groups discovered in this network. The groups for the most part follow conference membership. Cases where deviations from conference membership occur have natural interpretations, such as geographic proximity of teams. In some cases, the groupings reflect past associations, with a team being assigned to a group with other teams from its former conference, rather than the new conference it belongs to. More interestingly, we find that deviations from conference membership predict future developments, specifically, teams switching conference membership after 2001. For example, New Mexico State (Sun-Belt Conference) was grouped with Western Athletic (WAC) by our algorithm. It joined WAC in 2005. Texas Christian (Conference USA) was also grouped with WAC. It was part of WAC but joined Conference USA in 2001. Central Florida

(Independent) was grouped with Mid-American conference by the algorithm, and joined it in 2002. Notre Dame (Independent) was grouped with Big 10 conference, and as of 2008, is in talks of joining it. Alternatively, we can represent the College Football dataset as a unipartite graph, where the vertices are teams and edges represent regular season game between the teams [21]. Table II(a) shows the mutual information-based measure of the quality of the discovered groups vs α in this network. Note that maximum value of α is bigger for this network. The communities are less pure than those discovered using the 2-mode matrix. Overall, the heterogeneous method gives a crisper division. When it does put conferences together, these assignments make sense for geographic or historic reasons, and as we showed above, sometimes anticipate future developments.

C. Flickr Social Network

We also ran our algorithm on the heterogeneous social network data collected from Flickr, a social photosharing site that allows users to upload images, post them to special interest photo groups, and to join social networks by adding other users as friends or contacts. Since the actual social network on Flickr is rather vast, we sampled it by identifying users who were broadly interested in one of three topics [25]: child and family *portraiture*, *nature* photography and *technology*. For each topic, we used the Flickr API to perform a tag search using a keyword relevant to that topic, to retrieve 500 ‘most interesting’ images. We then extracted the names of users who submitted these images to Flickr and added them to our data set. The keywords used for image search were (a) *newborn* for the *portraiture* topic, (b) *tiger* and *beetle* for the *nature* topic, and (c) *apple* for the *technology* topic. Each keyword is ambiguous. *Tiger*, for example, could mean a wild animal, but also a flower (Tiger lily), Mac operating system (OS X Tiger), or a famous golfer (Tiger Woods), while *beetle* could describe a bug or a car.

From the set of users in each topic, we identified four (eight for *nature*) who were interested in each topic. We examined each user’s profile to confirm that the user was indeed interested in that topic. Specifically, we looked at group membership and user’s most common tags. Thus, groups such as “Big Cats”, “Zoo”, “The Wildlife Photography”, etc. pointed to user’s interest in the *nature* topic. We used the Flickr API to retrieve the contacts of these users, as well as their contacts’ contacts. We labeled contacts by the topic of the seed user. Although we did not verify that all the labeled users were indeed interested in the topic, we use these *soft labels* to evaluate the discovered communities.

1) *Communities*: Once we retrieved the social networks of a target set of users, we reduced it to an undirected network containing mutual contacts only. In other words, every link in the network between two nodes, say *A* and *B*, implies that *A* lists *B* as contact and *vice versa*. This resulted in a network of 5747 users. Of these, 1620 users were labeled *technology*, 1337 and 2790 users were labeled *portraiture* and *wildlife* respectively. The normalized mutual information for

community division of this network is shown in Table II(c). We took the soft labels corresponding to topic of photography interest as the true community division of the network. As α increases up to its maximum value, the groups become purer, and the mutual information increases. Except for the maximum value of α , there were three groups. Group1 was composed mainly of *technology* users, Group2 mainly of *wildlife* users. Users interested in *portraiture* emerged as a distinct group, Group3, whose size was largely independent of α . The fourth group found at $\alpha = 0$ was a mixture of all topics, and at the maximum value of α only two groups were found.

Next, we augmented the mutual contacts data with information about user membership in public groups on Flickr. We used the Flickr API to retrieve the public groups to which the users in our dataset belonged, 51,000 groups in total. We considered a user to be active in a group if among the most recent 100 photos she uploaded, more than 10 were posted to that group. We were able to extract the active groups for 3625 of the 5747 users, and they belonged to 10,463 active groups.

We represented this data as a 2-mode matrix of users and groups, where a relation between users specified whether they were each other’s mutual contacts, and a relation between a user and a group specified whether the user was an active member of this group. The 2-mode matrix proved too large for eigenvector decomposition on the computing resources available to us. Instead, we used the Lanczos algorithm [26] to efficiently compute the leading eigenvalue of the 2-mode matrix and then used the eigenvector corresponding to this eigenvalue to optimize modularity. Table II(d) evaluates the quality of the community division of the heterogeneous network. While the normalized mutual information metric is worse than for the mutual contacts network, looking closer at the results suggests that the heterogeneous network has a somewhat different structure. In the mutual contacts network, the *portraiture* group emerged as a distinct group, probably because its members are tightly interconnected. In the heterogeneous user-group network, the *nature* group emerges as a distinct group. Although members of this group seem to be less well-connected as contacts, they appear to be active in similar groups. Our method takes user-group relations into account and is able to identify these users to yield new insights into the structure of the Flickr community.

2) *Rankings*: We ranked users according to their b-centrality. Unfortunately, there is no independent analysis of the rankings of users, so we do not have a gold standard to evaluate the results of our algorithm. Figure 2 shows how the rankings of users relative to their ranking at $\alpha = 0$ change with increasing α in the (a) mutual contacts and (b) user-group networks. We claim that nodes whose rank improves with α (5, 10, 14, 16, 19) are the bridging nodes. Though we have no way to confirm it, it appears that these users appeal to others outside their community. Other nodes (4, 8, 18) see their rank worsen with α . These are nodes peripheral to the *portraiture* group that dominates the rankings. The mutual contacts-based rankings correlate somewhat with PageRank-based rankings.

Rankings of the user-group network (Fig. 2(b)) produce

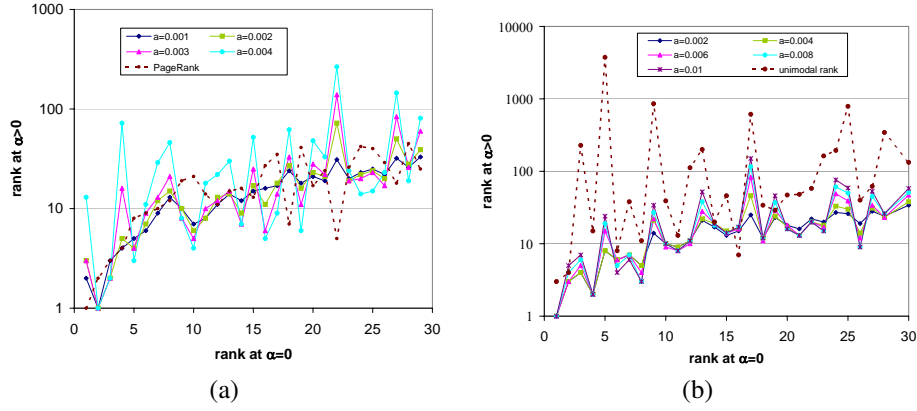


Fig. 2. Rankings of select Flickr users in the (a) mutual contacts and (b) user-group networks

similar trends, though there are only 9 users who were in the top-ranked set shown in Fig. 2(a). The new top-ranked users are mostly from the *nature* group. Although the data is difficult to evaluate, taking user-group relations into account appears to emphasize the importance of the *nature* group.

V. RELATED WORK

Liben-Nowell and Kleinberg [18] have shown that Katz measure is the most effective measure for the link prediction task, better than hitting time, PageRank [20] and its variants. Unlike the Katz score, Bonacich centrality [4], remained relatively unknown in the computer science community. It parametrizes the Katz score with α , a parameter that gives the weight of distant links, and also sets the scale of the centrality measure. We showed the benefit of using this parameter in the analysis of network structure.

There has been some work in motif-based communities in complex networks [27] which like our work extends traditional notion of modularity introduced by Girvan and Newman [21]. The underlying motivation for motif-based community detection is that “the high density of edges within a community determines correlations between nodes going beyond nearest-neighbours,” which is also our motivation for applying centrality-based modularity to community detection. Though the motivation of this method is to determine the correlations between nodes beyond nearest neighbors, yet it does impose a limit on the proximity of neighbors to be taken into consideration dependent on the size of the motifs. The method we propose, on the other hand, imposes no such limit on proximity. On the contrary, it considers the correlation between nodes in a more global sense. The measure of global correlation evaluated using the b-centrality metric would be equal to the weighted average of correlations when motifs of different sizes are taken. B-centrality enables us to calculate this complex term quickly and efficiently.

Resolution limit is one of the main limitations of the original modularity detection approach [28]. It can account for the comment by Leskovec et al. [29] that they “observe tight but almost trivial communities at very small scales, the best possible communities gradually ‘blend in’ with rest of the network

and thus become less ‘community-like.’” However, that study is based on the hypothesis that communities have “more and/or better-connected ‘internal edges’ connecting members of the set than ‘cut edges’ connecting to the rest of the world.” Hence, like most graph partitioning and modularity-based approaches to community detection, their process depends on the local property of connectivity of nodes to neighbors via edges and is not dependent on the structure of the network on the whole. Therefore, it does not take into account connectivity in a more global sense, as given by centrality metrics. In their paper on motif-based community detection, Arenas et al. [27] state that the extended quality functions for-motif based modularity also obey the principle of the resolution limit. But this limit is now motif-dependent and then several resolution of substructures can be achieved by changing the motif. However, it would be difficult to verify which resolution of substructures is closest to natural communities. In b-centrality-based modularity, on the other hand, the resolution limit depends on the centrality radius, given by the attenuation factor α . Smaller α lead to smaller radii, and, therefore, to division of the network into a larger number of communities [12].

There have been two recent works that extend modularity-based approach to bipartite networks [1], [2]. Both of these methods express modularity in terms of edges; therefore, their formulation of modularity maximization suffers from the same problem of localization as the original formulation by Newman for unipartite graphs, and are unable to determine correlation between nodes beyond nearest neighbors. We, on the other hand, can vary parameter α to take nodes beyond nearest neighbors into account. Barber et al. [2] argue that in the representation of modularity of used by Guimera et al. [1] identifies modules in only one part of the network at a time. They, on the other hand, classify nodes in both partitions simultaneously and customize spectral methods to bipartite graphs. The customization is based on the identification of the asymmetric submatrix of the full bipartite modularity matrix. This asymmetric submatrix not only represents the bipartite nature of the graph, but also enables them to customize bipartite modularity-maximization method by using

singular value decomposition and recursive identification of bipartite modules. However since this algorithm explicitly takes advantage of the bipartite nature of the graph, it cannot be used for graphs containing intra-layer edges along with inter-layer edges. For example, in the case of Flickr, bipartite representation may capture only user-group relations, but not information about user-user or group-group relations. Hence our method is more appropriate to capturing the complete information encoded within a social network.

VI. CONCLUSIONS

In this paper, we introduced a compact data structure, the N-mode matrix, to represent different classes of entities and relations present in a heterogeneous network. We used Bonacich centrality to study the structure of such networks, specifically, identify communities and important nodes in the network. We extended the modularity optimization-based class of algorithms to use b-centrality, rather than edges, as a measure of network connectivity. We applied this approach to benchmark networks studied in literature and found that it results in network division in close agreement with the ground truth. In addition, it gave useful insights into the structure of the graph and information about the changes that happen in the future, but were not known at the time when data was collected. We also used b-centrality to rank nodes in a network. By studying changes in rankings that occur when the indirect attenuation factor α changes, we were able to identify leaders and ‘bridging’ nodes that facilitate communication between different communities. The results of the community-finding algorithm applied to Flickr network were mixed. One possibility is that since the number of groups and group membership is much larger than the number of users, group information completely masks user-user information. We may want to differentially weigh relations to balance transmission of influence along different channels. To do this, we break the 2-mode matrix into diagonal (intra-layer) and off-diagonal (inter-layer) components: $A = D_1 A_1 + D_2 A_2$, where

$$A_1 = \begin{bmatrix} XX_{m \times m} & 0_{m \times n} \\ 0_{n \times m} & YY_{n \times n} \end{bmatrix} \quad A_2 = \begin{bmatrix} 0_{m \times m} & XY_{m \times n} \\ YX_{n \times m} & 0_{n \times n} \end{bmatrix},$$

and weights are given by matrices

$$D_1 = \begin{bmatrix} D_{m \times m}^\gamma & 0_{m \times n} \\ 0_{n \times m} & D_{n \times n}^\delta \end{bmatrix} \quad D_2 = \begin{bmatrix} D_{m \times m}^\lambda & 0_{m \times n} \\ 0_{n \times m} & D_{n \times n}^\mu \end{bmatrix},$$

with each D^γ, \dots, D^μ a diagonal matrix with $D_{ii}^\gamma = \gamma$, etc. We plan to study this balancing scheme on real-world networks.

ACKNOWLEDGMENT

This work is based in part on research supported by the National Science Foundation under awards BCS-0527725 and CMMI-0753124.

REFERENCES

- [1] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, “Module identification in bipartite and directed networks,” *Physical Review E*, vol. 76, no. 3, 036102, 2007.
- [2] M. J. Barber, “Modularity and community detection in bipartite networks,” *Physical Review E* 76, vol. 6, 2007.

- [3] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical Review E*, vol. 69, 026113, 2004.
- [4] P. Bonacich, “Power and centrality: A family of measures,” *The American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.
- [5] H. Tong, S. Papadimitriou, P. S. Yu, and C. Faloutsos, “Proximity tracking on time-evolving bipartite graphs,” *SDM. SIAM*, 704–715, 2008.
- [6] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika*, vol. 18, pp. 39–40, 1953.
- [7] L. C. Freeman, “Centrality in social networks: Conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.
- [8] W. L. Ferrar, *Finite Matrices*. Oxford Univ. Press, 1951.
- [9] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E*, vol. 69, p. 066133, 2004.
- [10] —, “Detecting community structure in networks,” *The European Physical Journal B*, vol. 38, pp. 321–330, 2004.
- [11] —, “Finding community structure in networks using the eigenvectors of matrices,” *Physical Review E*, vol. 74, p. 036104, 2006.
- [12] R. Ghosh and K. Lerman, “Community detection using a measure of global influence,” in *Proc. of the 2nd KDD Workshop on Social Network Analysis (SNAKDD’08)*, 2008.
- [13] U. Brandes, D. Dellling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On modularity clustering,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 20, no. 2, pp. 172–188, 2008.
- [14] E. A. Leicht and M. E. J. Newman, “Community structure in directed networks,” *Physical Review Letters*, vol. 100, p. 118703, 2008.
- [15] K. C. Foster, S. Q. Muth, J. J. Potterat, and R. B. Rothenberg, “A faster katz status score algorithm,” *Comp. Math. Organ. Theory*, 7, 4, 275–285, 2001.
- [16] G. Simmel, *The Sociology of Georg Simmel*. Free Press, 1950, ch. Individual and Society.
- [17] R. S. Burt, *Structural Holes: The Structure of Competition*. Cambridge, MA: Harvard University Press, 1992.
- [18] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [19] M. Granovetter, “The strength of weak ties,” *The American Journal of Sociology*, May 1973.
- [20] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web,” Stanford Digital Library Technologies Project, Tech. Rep., 1998.
- [21] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proc. Natl. Acad. Sci. USA*, vol. 99, p. 7821, 2002.
- [22] L. Freeman, “Finding social groups: A meta-analysis of the southern woman data,” *Dynamic Social Network Modeling and Analysis*, 2002.
- [23] L. Danon, A. Daz-Guilera, J. Duch, and A. Arenas, “Comparing community structure identification,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 9, pp. P09 008–09 008, 2005.
- [24] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South*. Chicago: The University of Chicago Press, 1941.
- [25] K. Lerman, A. Plangprasopchok, and C. Wong, “Personalizing results of image search on flickr,” in *AAAI workshop on Intelligent Techniques for Web Personalization*, 2007.
- [26] C. Lanczos, “An iteration method for the solution of the eigenvalue problem of linear differential and integral operators,” *Journal of Research of the National Bureau of Standards*, vol. 45, no. 4, Oct. 1950.
- [27] A. Arenas, A. Fernandez, S. Fortunato, and S. Gomez, “Motif-based communities in complex networks,” *Mathematical Systems Theory*, vol. 41, p. 224001, 2008.
- [28] S. Fortunato and M. Barthelemy, “Resolution limit in community detection,” *PROC. NATL. ACAD. SCI. USA*, vol. 104, p. 36, 2007.
- [29] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, “Statistical properties of community structure in large social and information networks,” in *Proceedings of the World Wide Web Conference*, 2008.
- [30] R. L. Breiger, “The duality of persons and groups” in *Social Forces*, vol. 53, p. 181–190, 1974.
- [31] L. C. Freeman and D. R. White, “Using Galois lattices to represent network data” in *Sociological Methodology 1993*, P. Marsen, ed. p. 127–146, 1994.